# **Chapitre 6** : Introduction aux modèles linéaires généralisés

Dans tout ce qui a précédé, on a cherché à prédire

- une réponse $Y$ **numérique**,
- à l'aide de covariables $X_1, \ldots, X_p$ **numériques**.

Et on a vu comment revenir à des covariables numériques si initialement catégorielles.

Dans ce chapitre, on va s'intéresser à deux autres cas :

(i) $Y \in \{0, 1\}$ est **binaire** (Exemples : sain/malade ; remboursera/fera défaut ; républicain/démocrate, etc.)

(ii) $Y \in \mathbb{N}$ est un **comptage** (Exemples : nombres d'enfants ; nombres de décès causés par... ; nombres de buts dans un match)

Dans ces deux cas, l'hypothèse du **modèle linéaire gaussien**, à savoir

$$[Y|X_{1:p}] \sim \mathscr{N}\big(\mu(X_{1:p}), \sigma^2\big),$$

n'est pas réaliste. Il faut donc en changer. En outre, on rappelle que, dans le modèle linéaire gaussien, on a

$$\mathbb{E}(Y|X_{1:p}) = \mu(X_{1:p}) = \beta_0 + \sum_j \beta_j X_j.$$

On va donc regarder deux modèles.

(i) Si $Y \in \{0,1\}$ est **binaire**, il faut supposer que

$$[Y|X_{1:p}] \sim \mathscr{B}\big(\mu(X_{1:p})\big).$$

(ii) si $Y \in \mathbb{N}$ est un **comptage**, on va supposer que

$$[Y|X_{1:p}] \sim \mathscr{P}\big(\mu(X_{1:p})\big).$$

Dans les deux cas,

$$\mathbb{E}(Y|X_{1:p}) = \mu(X_{1:p})$$

mais on ne va pas supposer que cette espérance conditionnelle est de la forme $\beta_0 + \sum_j \beta_j X_j$ car il y a des **contraintes** :

(i) $0 \leq \mu(X_{1:p}) \leq 1$ dans le cas de la loi de Bernoulli ou

(ii) $\mu(X_{1:p}) \geq 0$ dans le cas de la loi de Poisson.

# 1 Régression logistique

On s'intéresse ici au premier cas, où $Y \in \{0,1\}$ est **binaire**, et on suppose que

$$[Y|X_{1:p}] \sim \mathscr{B}\big(\mu(X_{1:p})\big).$$

Dans ce cas, $\quad \mu(X_{1:p}) = \mathbb{E}(Y|X_{1:p}) = \mathbb{P}(Y=1|X_{1:p}) \in [0;1].$

**La fonction logistique** est définie, pour tout $x \in ]0;1[$ par

$$\begin{aligned} \text{logit} : \ ]0;1[ &\to \mathbb{R} \\ x &\mapsto \log\left(\frac{x}{1-x}\right). \end{aligned}$$

C'est une bijection strictement croissante de $]0;1[$ sur $\mathbb{R}$, infiniment dérivable. Et la fonction inverse est donnée par

$$\begin{aligned} \text{logit}^{-1} : \ \mathbb{R} &\to ]0;1[ \\ x &\mapsto \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}} \end{aligned}$$

qui est infiniment dérivable. Voici le graphe de cette fonction.

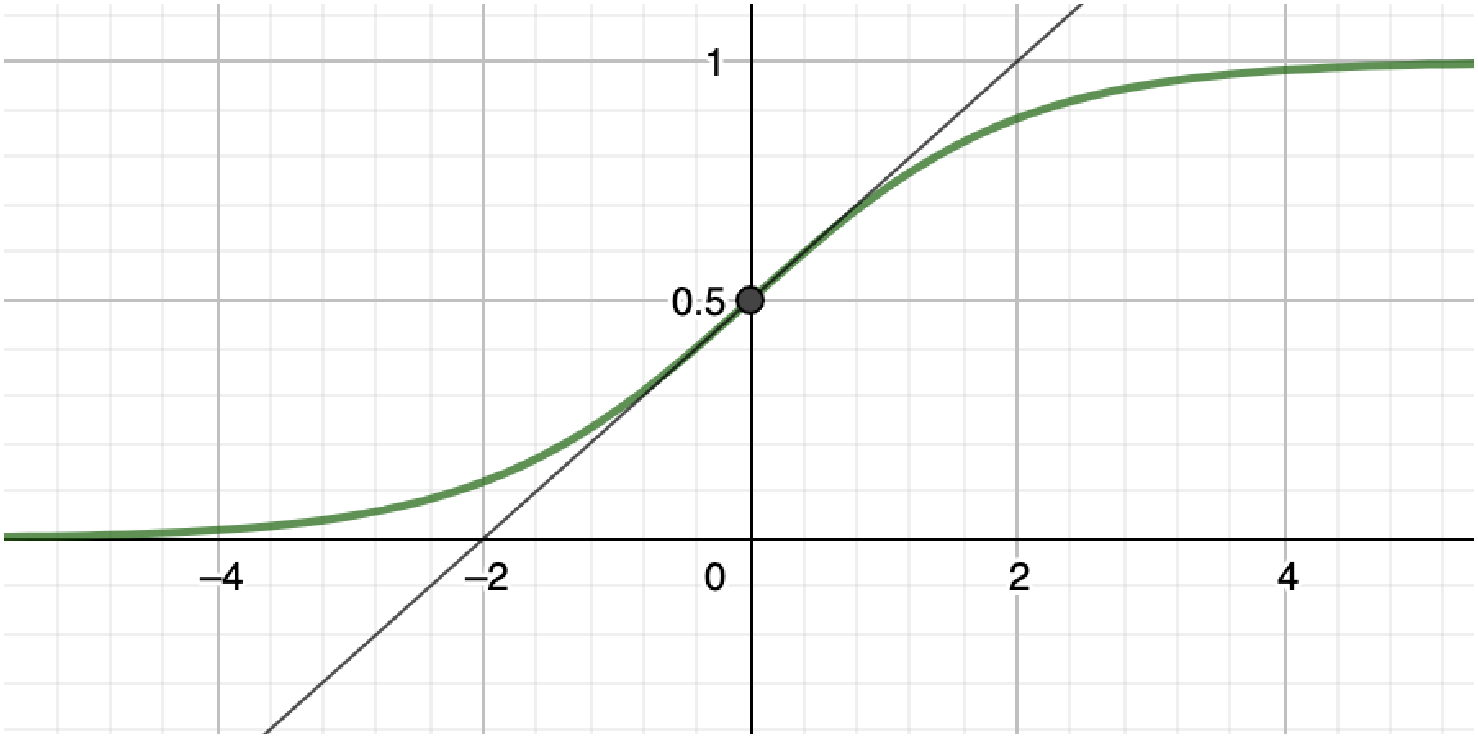


Figure 1 – *Graphe de la fonction logit*$^{-1}$*, et sa tangente en* $x = 0$

L'**hypothèse linéaire** devient maintenant

$$\text{logit}\big(\mu(X_{1:p})\big) = \beta_0 + \sum_j \beta_j X_j.$$

Autrement dit,

$$\begin{aligned}\mathbb{E}(Y|X_{1:p}) = \mathbb{P}(Y=1|X_{1:p}) &= \mathrm{logit}^{-1}\Big(\beta_0 + \sum_j \beta_j X_j\Big) \\ &= \frac{1}{1+\exp\Big(-\beta_0 - \sum_j \beta_j X_j\Big)} \in ]0;1[.\end{aligned}$$

Le **rapport de cote** (ou *odds ratio*) vaut alors

$$\frac{\mathbb{P}(Y=1|X_{1:p})}{\mathbb{P}(Y=0|X_{1:p})} = \exp(\beta_0)\prod_{j=1}^{p}\exp(\beta_j X_j).$$

Ce qui revient à supposer un **effet multiplicatif** des covariables sur le rapport de cote. En effet, toute autre covariable étant fixée, remplacer $X_j$ par $X_j+1$ revient à multipier le rapport de cote par $\exp(\beta_j)$.

Avec des données $\mathbf{Y} \in \{0,1\}^n$ et $\mathbf{X} \in \mathbb{R}^{n\times p}$, on ajuste le modèle par **maximum de vraisemblance**. Ce maximum est n'est pas explicite, il faut utiliser un algorithme d'optimisation numérique pour le trouver (Newton-Raphson). Ici, la vraisemblance est

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \quad \text{où } p_i = \frac{1}{1+\exp(-\beta_0 - \sum_j \beta_j X_{ij})}.$$

On note $\widehat{\boldsymbol{\beta}}$ l'estimateur du maximum de vraisemblance.

Pour un **nouvel individu** de covariables $\widetilde{X}_1, \dots, \widetilde{X}_p$ connues et de réponse $\widetilde{Y}$ inconnue, on peut **inférer/prédire** deux quantités différentes :

— la probabilité de la réponse $1$, ou la réponse moyenne, à savoir $\mu(\widetilde{X}_{1:p}) = \mathbb{P}\left(\widetilde{Y} = 1 \middle| \widetilde{X}_{1:p}\right)$, avec

$$\widehat{\mu}(\widetilde{X}_{1:p}) = \frac{1}{1 + \exp\left(-\widehat{\beta}_0 - \sum_j \widehat{\beta}_j \widetilde{X}_j\right)}$$

— la réponse $\widetilde{Y}$ elle-même, avec

$$\widehat{Y} = \begin{cases} 1 & \text{si } \widehat{\mathbb{P}}\left(\widetilde{Y} = 1 \middle| \widetilde{X}_{1:p}\right) > 0.5 \\ 0 & \text{sinon} \end{cases}$$

# 2 Régression de Poisson

**Rappel.** Si $Z \sim \mathscr{P}(a)$, alors,

$$\forall\ k \in \mathbb{N}, \quad \mathbb{P}(Z = k) = e^{-a}\frac{a^k}{k!} \quad \text{et} \quad \mathbb{E}(Z) = a.$$

## 2.1 Exposition constante

On s'intéresse ici au second cas, où $Y \in \mathbb{N}$ est un **comptage**, et on suppose que

$$[Y|X_{1:p}] \sim \mathscr{P}\big(\mu(X_{1:p})\big).$$

Dans ce cas, $\quad \mu(X_{1:p}) = \mathbb{E}(Y|X_{1:p}) \in [0; +\infty[.$

Dans la régression de Poisson, la fonction **logarithme** va jouer le même rôle que la fonction logistique et on suppose que

$$\log \mu(X_{1:p}) = \log \mathbb{E}(Y|X_{1:p}) = \beta_0 + \sum_j \beta_j X_j.$$

Avec des données $\mathbf{Y} \in \mathbb{N}^n$ et $\mathbf{X} \in \mathbb{R}^{n\times p}$, on ajuste le modèle par **maximum de vraisemblance**. Ce maximum est n'est pas explicite, il faut utiliser un algorithme d'optimisation numérique pour le trouver (Newton-Raphson). Ici, la vraisemblance est

$$L(\beta) = \prod_{i=1}^{n} e^{-a_i} \frac{a_i^{y_i}}{y_i!}, \quad \text{où } a_i = \exp(\beta_0 + \sum_j \beta_j X_{ij}).$$

On note $\widehat{\boldsymbol{\beta}}$ l'estimateur du maximum de vraisemblance.

Pour un **nouvel individu** de covariables $\widetilde{X}_1, \dots, \widetilde{X}_p$ connues et de réponse $\widetilde{Y}$ inconnue, on peut **inférer/prédire** deux quantités différentes :

— la réponse moyenne, à savoir $\mu(\widetilde{X}_{1:p}) = \mathbb{E}\left(\widetilde{Y} \middle| \widetilde{X}_{1:p}\right)$, avec

$$\widehat{\mu}(\widetilde{X}_{1:p}) = \exp\left(\widehat{\beta}_0 + \sum_j \widehat{\beta}_j \widetilde{X}_j\right)$$

— la réponse $\widetilde{Y}$ elle-même, avec

$$\widehat{Y} = G\left(\widehat{\mu}(\widetilde{X}_{1:p})\right)$$

où, pour tout $a > 0$,

$$G(a) = \underset{k\in\mathbb{N}}{\operatorname{argmax}}\ e^{-a}\frac{a^k}{k!}.$$

Notons ici que $G(a)$ est la modalité la plus probable de la loi $\mathscr{P}(a)$.

## 2.2 Avec durée d'exposition variable

Une variable aléatoire de Poisson $Z$ peut être un comptage **pendant une durée donnée** $D > 0$. (Exemple : nombre d'accident pendant une certaine durée). On suppose alors que

$$\mathbb{E}(Z|D) = aD, \quad \text{où } a \text{ est une nombre moyen par unité de temps.}$$

Donc,

$$\log \mathbb{E}(Z|D) = \log a + \log D.$$

Dans le modèle de Poisson avec durée d'exposition variable, on va donc supposer que

$$\log \mathbb{E}(Y|D, X_{1:p}) = \beta_0 + \sum_j \beta_j X_j + \log D.$$

Alors,

— $E = \log D$ intervient dans la formule comme une covariable, mais le coefficient en facteur est connu, égal à $1$,

— $\exp\left(\beta_0 + \sum_j \beta_j X_j\right)$ est le nombre moyen par unité de temps (à covariables fixées).

L'ajustement se fait également par maximum de vraisemblance. La précision des estimateurs n'est pas liés au nombres d'observations, mais à la durée totale d'exposition

$$n_D = D_1 + D_2 + \ldots + D_n.$$

# 3 Exemples

## 3.1 Régression logistique simple

On s'intéresse à l'élection présidentielle américaine en 1992 qui opposait Georges W. Bush et Bill Clinton (vainqueur). S'ajoute un troisième candidat sans étiquette Ross Perot. Pour étudier le vote d'un électeur, on pose

$$Y = \begin{cases} 1 & \text{si vote en faveur de Bush} \\ 0 & \text{sinon} \end{cases}$$

On dispose d'une variable explicative $X \in \{1, 2, \ldots, 5\}$ qui indique le niveau de revenu de l'électeur. (Pauvre $\iff X = 1$). Données : 14031 votants.

Table 1 – *Fréquence des votes par classes de revenus ($Y = 1$ si vote Bush)*

| **income** | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $\%(Y=1)$ | 41 | 44 | 49 | 54 | 70 |
| $\%(Y=0)$ | 59 | 56 | 51 | 46 | 30 |

Si on ajuste un modèle de régression logistique sur ces données, on obtient

$$\text{logit}\Big(\mathbb{P}(Y=1|X)\Big) \approx -0.67 + 0.23X.$$

On peut représenter l'ajustement avec la figure ci-dessous.

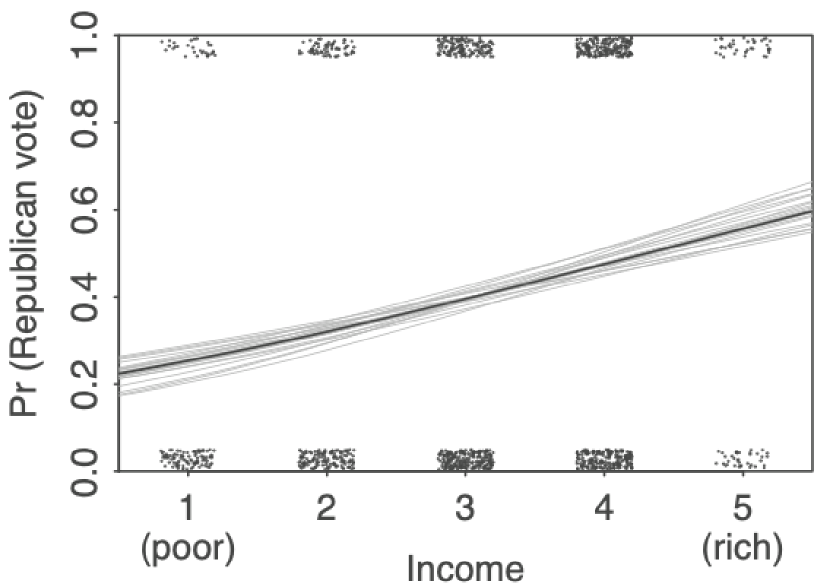


Figure 2 – Ajustement sur l'élection de 1992

Comment **s'interprète** $0.23$ ? Quand on remplace income $= x$ par income $= x+1$, le **rapport de cote** est multiplié par

$$\exp(0.23) \approx 1.25.$$

Et les prédictions sont

Table 2 – *Fréquence des votes par classes de revenus ($Y = 1$ si vote Bush)*

| **income** | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **données** $\%(Y = 1)$ | 41 | 44 | 49 | 54 | 70 |
| **prédictions** $\widehat{\mu}$ | 0.39 | 0.44 | 0.50 | 0.56 | 0.62 |

## 3.2 Nombre d'accidents de la route

On souhaite prédire le nombre d'accidents de la route aux carrefours par année en fonction de deux covariables :

— $X_1$, vitesse moyenne en km/h des voitures sur les routes autour du carrefour,

— $X_2$, présence de feu tricolore au carrefour.

Avec des données, on a ajusté le modèle

$$\log \mathbb{E}\left(Y \middle| X_1, X_2, D\right) \approx 2.8 + 0.02 X_1 - 0.20 X_2 + \log D.$$

**Interprétations :**

— Si on augmente la vitesse moyenne de 10 km/h sans changer $X_2$, le nombre moyen d'accidents par années est multiplié par $\exp(0.02 \times 10) \approx 1.22$. D'où une augmentation du nombre moyen d'environ $22\%$.

— Si on ajoute un feu tricolore sans changer $X_1$, le nombre moyen d'accidents par années est multiplié par $\exp(-0.2) \approx 0.82$[illegible]. D'où une diminution de $18\%$ du nombre moyen d'accidents.